International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Performance Analysis of Data Mining Algorithms in the Prediction of Rain Fall

Yogesh Kumar Jakhar, Nidhi Mishra, Rakesh Poonia

Abstract*: Weather being a random phenomenon its prediction has been always a challenge for the meteorologist all over the world. There are number of approaches for predicting this weather based on atmospheric data collected. Rain forecasting is a puzzling, composite, vigorous and mind-boggling task. Rain forecasting pretenses right from the primeval times as a challenging task, because it be influenced by numerous parameters like temperature, wind speed and direction, rainfall, humidity, station level pressure, mean sea-level pressure, dry bulb temperature, dew point temperature and vapour pressure. Various data mining techniques were implemented for rain forecasting. With compared to orthodox methods predicting rainfall rate, the methods that were applying chronological records and data mining technology shows improvement in computing accurate results with more accuracy. Many researchers have done excellent works to construct forecasting models with data mining methods;but in them most just test the predicting accuracy at one particular geographical area. In this paper, we analyzed the performance of k-NN, Random Forest, C5.0 and AdaBoost algorithms on different locations and compared the performance using precision, recall, f-measure and classification accuracy. The daily surface data was collected from India Meteorological Department (IMD), Pune of 3 stations form the period 2005 to 2015. The k-NN algorithm perform better accuracy 98.02 % on Jodhpur dataset with compare to other datasets, the ratio of 90:10 of training and testing records and the value of K is 10. The highest accuracy is 99.270 % of AdaBoost algorithm.

Index Terms*: AdaBoost, C5.0, k-NN, , Rainfall prediction, Random Forest, and Weather forecasting.

I. INTRODUCTION

Precipitation is a complex atmospheric process, which relies on many climate related highlights. Precise and convenient precipitation forecast can be useful in numerous courses, for example, arranging the water assets the executives, issuance of early surge alerts, dealing with the flight tasks and restricting the vehicle and development exercises [2]. Exact precipitation expectation is increasingly perplexing today because of atmosphere varieties. Researchers reliably have been attempting to foresee precipitation with most extreme precision by upgrading and incorporating information mining methods [4]. A wide assortment of precipitation conjectures techniques are accessible in India since India is an agrarian nation and the achievement of farming depends of precipitation. There are for the most part two ways to deal with anticipating precipitation. They are Empirical strategy and dynamical technique. The experimental methodology depends on investigation of authentic information of the precipitation and its relationship to an assortment of barometric factors. The most broadly utilized observational methodologies utilized for atmosphere expectation are relapse, Artificial Neural Network(ANN), fluffy rationale and bunch strategy for information taking care of. In dynamical methodology, forecasts are created by physical models dependent on frameworks of conditions that foresee the advancement of the worldwide atmosphere framework because of starting climatic conditions [1]. The paper is organized as follows. Related work to data mining algorithms for rain prediction is presented in section 2. Section 3 describes about dataset, parameters description and experiment evaluation process. The experimental results are shown and discussed in section 4. Section 5 presents conclusions about performance analysis of k-NN, Random Forest, C5.0 and AdaBoost algorithms using R Programming.

II. LITERATURE REVIEW

In this subdivision, we concisely review some back ground research. Amira Shoukry and Ahmed Rafea discovered the effect of corpus size and features on the machine leaning classification quality. They proposed a system which showed better results than other sentence-level sentiment grouping systems [1]. Gino Slanzi, Jorge Balazs and Juan D. Velasquez studied user's behavior when interacting on internet with websites from a physiological point of view. They predicted clicks intention, using pupil dilation and EEG responses with experiment where ET and EEG of 21 fit persons. Seven methods were experienced with 15out 789 pupil dilation and EEG features acquired from a Random Lasso selection process, showing good results for Accuracy with 71.09% using Logistic Regression [2]. Keita Mori and Osamu Mizuno applied a prototype fault-prone module forecasting tool using a text-filtering based technique named "Fault Prone Filtering". It experimented on three open source projects and the results shows that tool detect 90 % of the actual fault modules with the accuracy of 67 % [3]. Timo Johann, Christoph Stanik, Alireza M. Alizadeh B. and Walid Maalej designed a simple approach for feature extraction (SAFE) from app pages and user reviews.

Revised Manuscript Received on 30 July 2019.

* Correspondence Author

Yogesh Kumar Jakhar*, Ph. D. Scholar, Department of Computer Engineering, Poornima University, Jaipur (India)

Dr. Nidhi Mishra, Associate Professor, Poornima University, Jaipur (India)

Dr. Rakesh Poonia, Assistant Professor, Govt. Engineering College, Bikaner (India)

The proposed SAFE approach performs a complete preprocessing of the natural language input, applies Part-of Speech and sentence patterns to extract human readable. The results on app descriptions were obtained an accuracy of up to 88% with an average of 56% and a recall of 70% with an average of 43%. The results for unfiltered user reviews obtained with an average precision of 24% but achieved a recall of 71% [4].

Moataz H. Khalil, Walaa M. Sheta, Adel S. Elmaghra by presented a categorizing technique to identify machines removed from the system based on failure or due to maintenance. The researcher also developed a support vector machine (SVM)model for learning and forecast of machine failure. The regionalized model achieved 99.04 % accuracy [6]. Qian Liu, Hui Jiao and Hui-bo Jia presented a word-based Chinese document experimental system which was aimed to make Chinese information processing technology more reliable and efficient. The Accuracy ratio P was 95.03%, recall ratio R was 91.40% and F-value was 93.18% [7]. Lichao Sun, Zhiqiang Li, Qiben Yan, Witawas Srisa-an and Yu Pan designed SIGPID, it was a malware detection system and it was based on permission analysis to cope with the rapid increase in the number of Android malware. The results show 90% of accuracy when Support Vector Machine (SVM) was used [10].

III. EXPERIMENT EVALUATION

This section describes the performance of k-NN, Random Forest, C5.0 and AdaBoost algorithms. The descriptions of the data sets and performance measures are given below.

A. Description of Data Sets

For experimental work, data were collected from India Meteorological Department (IMD), Pune. The required daily surface data of 3 stations were collected form the period 2005 to 2015 as available in the archives. These stations are Churu (42170), Jaipur (42348) and Jodhpur (42339). The data are in .CSV file. The dataset has 24 parameters. Following table have more details about parameters.

Table- 1: Parameter description

S. No.	Name of Parameter	Description
1	SLP	STATION LEVEL PRESSURE (in hpa)
2	MSLP	MEAN SEA-LEVEL PRESSURE (in hpa)
3	DBT	DRY BULB TEMPERATURE (in deg. C)
4	WBT	WET BULB TEMPERATURE (in deg. C)
5	DPT	DEW POINT TEMPERATURE (in deg. C)
6	RH	RELATIVE HUMIDITY (in %)
7	VP	VAPOUR PRESSURE (in hpa)
8	DD	WIND DIRECTION
9	FFF	WIND SPEED (in Kmph)
10	Cl	FORM OF LOW CLOUD
11	A	AMOUNT OF LOW CLOUD (in oktas)
12	Cm	FORM OF MEDIUM CLOUD (in codes)
13	A	AMOUNT OF MEDIUM CLOUD (in oktas)
14	Ch	FORM OF HIGH CLOUD (in code)
15	A	AMOUNT OF HIGH CLOUD (in oktas)
16	Dl	DIRECTION OF LOW CLOUD
17	Dm	DIRECTION OF MEDIUM CLOUD (in 8 points of compass)
18	Dh	DIRECTION OF HIGH CLOUD (in 8 points of compass)
19	TC	TOTAL AMOUNT OF CLOUDS (in oktas)
20	h	HEIGHT OF LOWEST CLOUD (in code)
21	c	FORM OF INDIVIDUAL LAYER OF CLOUD (in code)
22	a	AMOUNT OF INDIVIDUAL LAYER OF CLOUD (in oktas)
23	Ht	HEIGHT OF BASE INDIVIDUAL CLOUD LAYER (in code)
24	Rain	Rain fall

B. Performance Measurement

To calculate the efficiency of algorithms, there are numerous catalogs, like specificity, sensitivity, precision, and accuracy to measure the models' strength. These indices (equation 1-4) are deliberate by the confusion matrix (Table-2). We used the standard recall, precision, accuracy and f-measure aggregated over all classes. These are defined as [8]:

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (1)$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (2)$$

$$accuracy = \frac{\#\ of\ correctly\ classified\ documents}{total\ number\ of\ documents} \quad (3)$$

$$f-measure = \frac{2\ \times recall\ \times precision}{recall + precision} \quad (4)$$

Confusion matrix is a valuable tool for evaluating the performance of cataloging method in data mining. In the confusion matrix relevant data with the observations should be positioned on the main diagonal of the matrix, and the remaining values of the matrix are zero or near zero [12].

Table- 2: Confusion matrix

		Class	
		Negative (N)	Positive (P)
Predicted Class	Negative (-)	True Negative Count (TN)	False Negative Count (FN)

	Positive (+)	False Positive Count(FP)	True Positive Count (TP)

FN= Total number of positively categorized data, which have been categorized as "Negative" falsely.
TN= Total number of negatively categorized data, which have been categorized as "Correct".
TP= Total number of positively categorized data, which have been categorized as "Correct".
FP= Total number of negatively categorized data, which have been categorized as "Positive" falsely.

C. Experiment results

For experimental work the datasets are divided in two parts: Training dataset and testing dataset. We used 3 different types of training and testing records ration: these are 67:33, 80:20 and 90:10.

Predictive attributes or class: SLP, MSLP, DBT, WBT, DPT, RH, VP, DD, FFF, Cl, A, Cm, A, Ch, A, Dl, Dm, Dh, TC, H, C, a, Ht
Predicted attribute or class: Rain (Rain fall with 0 and 1 value)

C.1. k-NN Algorithm

The K- nearest neighbor method is a technique for classification based on closeness to different cases. Those near others are known as a "neighbor". At the point when a case is new, its separation from every one of the cases in the model is determined. Applying this characterization indicates the case just like the closest neighbor, which is the most comparable. Subsequently, it puts the case into the gathering that contains the closest neighbors. The calculation is additionally ready to ascertain esteems consistently for an objective. In this circumstance, the normal or the middle target estimation of the closest neighbor is utilized to get the anticipated estimation of new case. We applied k-NN algorithms on different K values. Different K values are: 1, 2, 3, 4, 5 and 10. The following table-3 has different ratios of training and testing datasets.

Table- 3: Accuracy of k-NN model on different K values

	Station								
	Churu			**Jaipur**			**Jodhpur**		
	Training and testing records ration (%)								
	67:33	**80:20**	**90:10**	**67:33**	**80:20**	**90:10**	**67:33**	**80:20**	**90:10**
K	**Accuracy of k-NN Model (%)**								
1	94.5	94.22	94.4	94.36	94.44	94.24	96.61	96.82	97.3
2	94.34	93.96	93.58	94.74	94.65	93.96	96.79	96.67	96.75
3	95.52	95.22	95.08	95.79	**96.09**	94.52	97.24	**97.71**	97.88
4	95.27	95.15	95.36	95.62	95.88	95.78	**97.29**	**97.71**	97.6
5	**96.45**	95.28	95.49	95.92	**96.09**	**96.34**	97.24	97.41	97.74
10	96.13	**95.48**	**95.77**	**96.09**	96.02	**96.34**	97.24	97.63	**98.02**

C.2 Random Forest algorithm

Random Forest is a versatile machine learning algorithm and it performs both regression and classification responsibilities. Random forest algorithm was applied on three different dataset with different ration of training and testing dataset. The following table-4 has confusion matrix for Random forest algorithm, experiment were performed on different ratios of testing and training datasets.

Table-4: Confusion Matrix and Statistics for Random Forest

Station		**Training and testing records ration (%)**					
		67:33		**80:20**		**90:10**	
	Rain Condition	No (0)	Yes (1)	No (0)	Yes (1)	No (0)	Yes (1)
Churu	No (0)	2357	85	1438	59	701	28
	Yes (1)	6	9	7	3	3	1
Ja	No (0)	2276	72	1391	51	670	26
	Yes (1)	9	22	5	12	1	7
Jodhpur	No (0)	2154	46	1323	23	637	18
	Yes (1)	3	13	2	6	1	1

C.3 C 5.0 Algorithm

C5.0 is an algorithm, developed by Ross Quinlan and it is used to construct a decision tree [11].C5.0 is a leeway of Quinlan's earlier ID3 algorithm. The C5.0 is frequently referred to as a numerical classifier. Writers of the Weka machine learning software defined the C5.0 algorithm as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date". The following table-5 has confusion matrix for C 5.0 algorithm, there are three different combinations of testing and training datasets.

Table-5: Confusion Matrix and Statistics for C5.0

Station	Rain Condition	67:33 No (0)	67:33 Yes (1)	80:20 No (0)	80:20 Yes (1)	90:10 No (0)	90:10 Yes (1)
Churu	No (0)	2362	1	1442	3	700	4
	Yes (1)	90	4	62	0	28	1
Jaipur	No (0)	2277	8	1392	4	668	3
	Yes (1)	70	24	50	13	24	9
Jodhpu	No (0)	2155	2	1317	8	639	0
	Yes (1)	48	11	22	7	19	0

Training and testing records ration (%)

C.4 AdaBoost Algorithm

AdaBoost, short for "Adaptive Boosting", is a machine learning algorithm designed by Yoav Freund and Robert Schapire. In general the AdaBoost model was used in combination with other learning algorithms to improve their performance. The following table-6 has confusion matrix for AdaBoost algorithm, there are three different combinations of testing and training datasets with 67:33, 80:20 and 90:10 ratios.

Table-6: Confusion Matrix and Statistics for AdaBoost

Station	Rain Condition	67:33 No (0)	67:33 Yes (1)	80:20 No (0)	80:20 Yes (1)	90:10 No (0)	90:10 Yes (1)
Churu	No (0)	4990	3	5906	5	6644	8
	Yes (1)	95	78	126	79	142	96
Jaipur	No (0)	4837	4	5722	8	6445	10
	Yes (1)	104	94	114	115	131	128
Jodhpu	No (0)	4555	2	5381	2	6046	1
	Yes (1)	31	68	38	85	49	93

Training and testing records ration (%)

IV. RESULTS AND DISCUSSION

The k-NN algorithm perform better accuracy 98.02 % on Jodhpur dataset with compare to other datasets, the ratio of 90:10 of training and testing records and the value of K is 10. For Churu data set got accuracy 95.77% with the ratio of 90:10 of training and testing records and the value of K is 10. For Jaipur data set got accuracy 96.34% with the ratio of 90:10 of training and testing records and the value of K are 5 and 10. In figure-1 it clearly shows that the accuracy of k-NN model higher for Jodhpur dataset in compare to Churu and Jaipur datasets in all ratios of training and testing records. Table-7 shows the resultant performance (recall, precision, accuracy and F-measure) of the k-NN, Random Forest, C5.0 and AdaBoost algorithms. It could be noticed that, the accuracy of AdaBoost is higher than k-NN, Random Forest and C5.0algorithms on all stations and all combinations of training and testing datasets ratios. With compare to previous work of many researcher they accuracy of AdaBoost (99.27%) is higher. For Jodhpur station the accuracy is high on the ratio of 80:20 of training and testing datasets

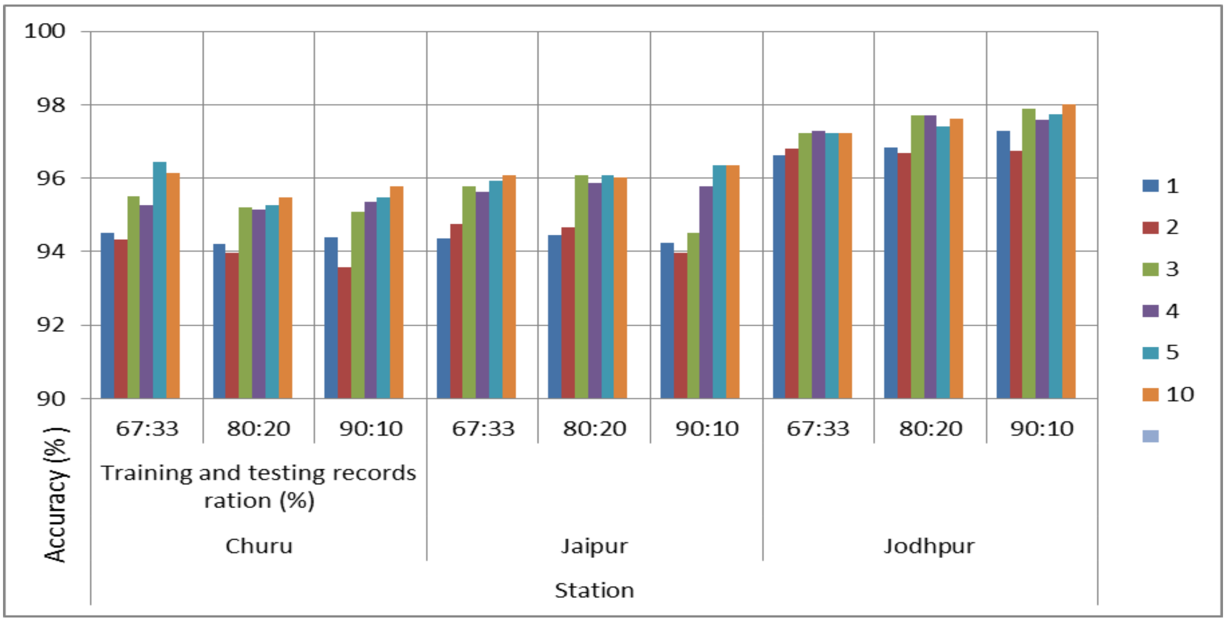

Fig.-1: Accuracy of k-NN model on different K values

Retrieval Number: B1604078219/19©BEIESP
DOI: 10.35940/ijrte.B1604.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table- 7: Evaluation measures using four different classifiers (k-NN, Random Forest, C5.0 and AdaBoost)

		Station								
		Churu			Jaipur			Jodhpur		
		Training and testing records ration (%)								
		67:33	80:20	90:10	67:33	80:20	90:10	67:33	80:20	90:10
k-NN	Recall	0.063	0.016	0.034	0.085	0.174	0.120	0.052	0.206	0.133
	Precision	0.462	0.125	0.250	0.533	0.688	0.429	0.250	0.429	0.667
	Accuracy (%)	96.130	95.480	95.770	96.090	96.090	96.350	97.290	97.710	98.030
	F-measure	0.112	0.029	0.061	0.147	0.278	0.188	0.087	0.279	0.222
Random Forest	Recall	0.600	0.300	0.250	0.709	0.705	0.875	0.812	0.750	0.500
	Precision	0.096	0.048	0.034	0.234	0.190	0.212	0.220	0.207	0.053
	Accuracy (%)	96.300	95.620	95.770	96.600	96.160	96.160	97.790	98.150	97.110
	F-measure	0.165	0.083	0.061	0.352	0.300	0.341	0.347	0.324	0.095
C 5.0	Recall	0.042	0.015	0.034	0.255	0.206	0.272	0.186	0.241	0.050
	Precision	0.800	0.250	0.200	0.750	0.765	0.750	0.840	0.460	0.500
	Accuracy (%)	96.300	95.690	95.630	96.720	96.300	96.160	97.740	97.780	96.970
	F-measure	0.081	0.030	0.059	0.381	0.325	0.400	0.306	0.318	0.091
AdaBoost	Recall	0.450	0.385	0.403	0.475	0.502	0.494	0.687	0.691	0.650
	Precision	0.963	0.940	0.923	0.959	0.935	0.928	0.970	0.977	0.989
	Accuracy (%)	**98.100**	**97.860**	**97.820**	**97.860**	**97.950**	**97.900**	**99.250**	**99.270**	**99.190**
	F-measure	0.612	0.547	0.561	0.635	0.635	0.645	0.805	0.810	0.788

V.CONCLUSION

Forecast is a puzzling task and that too for weather is even more complex, dynamic and mind-boggling. Weather forecast presents appropriate from the antiquated occasions as a major massive assignment, since it relies upon different parameters to foresee the needy factors like temperature, precipitation, dampness, wind speed and heading, which are changing occasionally and climate computation fluctuates with the geological area alongside its environmental factors. We used k-NN, Random Forest, C5.0 and AdaBoost algorithms. The performances of the classifiers are compared using precision, recall, f-measure and classification accuracy. The daily surface data was collected from India Meteorological Department (IMD), Pune of 3 stations form the period 2005 to 2015.For experimental workthe datasets are divided in two parts: Training dataset and testing dataset. We used 3 different types of training and testing records ration: these are 67:33, 80:20 and 90:10. The k-NN algorithm perform better accuracy 98.02 % on Jodhpur dataset with compare to other datasets, the ratio of 90:10 of training and testing records and the value of K is 10.The accuracy of AdaBoost is 99.270 % and it is higher than k-NN, Random Forest and C5.0algorithms on all stations and all combinations of training and testing datasets ratios.

REFERENCES

1. Amira Shoukryand Ahmed Rafea, "A Hybrid Approach for Sentiment Classification of Egyptian Dialect Tweets", 2015 First International Conference on Arabic Computational Linguistics, IEEE, pp. 78-85
2. Gino Slanzi, Jorge Balazs and Juan D. Velasquez, "Predicting Web user click intention using pupil dilation and electroencephalogram analysis", 2016 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 471-420
3. Keita Mori and Osamu Mizuno, "An Implementation of Just-In-Time Fault-PronePrediction Technique Using Text Classifier", 2015 IEEE 39th Annual International Computers, Software & Applications Conference, pp. 609-612
4. Timo Johann, Christoph Stanik, Alireza M. Alizadeh B. and Walid Maalej, "SAFE: A Simple Approach for Feature Extraction from App Descriptions and App Reviews", 2017 IEEE 25th International Requirements Engineering Conference, pp. 21-30.
5. Haifeng Liu, Zheng Hu, Dian Zhou and Hui Tian, "Cumulative Probability Distribution Model forEvaluating User Behavior Prediction Algorithms", SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013, pp. 385-391.
6. Moataz H. Khalil, Walaa M. Sheta and Adel S. Elmaghraby, "Categorizing Hardware Failure in Large Scale Cloud Computing Environment", 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT).
7. Qian Liu, Hui Jiao and Hui-boJia, "Study of Word-based Chinese Document Experimental System and Chinese Freetext Information Extraction Experiment Based on it", Third International Conference on Natural Computation (ICNC 2007), IEEE
8. Mohamed K. Elhadad, Khaled M. Badran, and Gouda I. Salama, "A Novel approach for Ontology-based Dimensionality Reduction for Web Text Document Classification", IEEE, 2017, pp. 373-378
9. Tanmay Basu and C. A. Murthy, "Effective Text Classification by a Supervised Feature Selection Approach", 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 918-925

10. Lichao Sun, Zhiqiang Li, Qiben Yan, WitawasSrisa-an and Yu Pan, "SigPID: Significant Permission Identification for Android Malware Detection",IEEE, 2016, pp. 59-66
11. Jinghao Niu and Wei Zhang, "Comparative Analysis of Statistical Models in RainfallPrediction", Proceeding of the 2015 IEEE International Conference on Information and Automation Lijiang, China, August 2015, pp. 2187-2190
12. Moloud Abdar, Sharareh R. Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto and Goli Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseses", International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 6, December 2015, pp. 1569-1576
13. Yogesh Kumar Jakhar, Nidhi Mishra and Rakesh Poonia, "Review on Weather Events Forecasting Using Data Mining Algorithms", Journal of Software Engineering Tools & Technology Trends, 2018, Volume 5, Issue 1, pp. 12-22
14. Mehmet Koyuturk, Ananth Grama, and Naren Ramakrishnan, "Compression, Clustering, and Pattern Discovery in Very High-Dimensional Discrete-Attribute Data Sets", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2005, VOL. 17, NO. 4, pp. 447-461
15. Mrutyunjaya Panda and Manas Ranjan Patra, "A COMPARATIVE STUDY OF DATA MINING ALGORITHMS FOR NETWORK INTRUSION DETECTION", First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008, pp. 504-507
16. Hidenao Abe and Shusaku Tsumoto, "Detecting Similarity of Transferring Datasets based on Features of Classification Rules", 2009 IEEE International Conference on Data Mining Workshops, pp. 412-415.
17. Shuo Wang and Xin Yao, "Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures", 2013, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, pp. 206-2019.

[18. Nasir Ali,Yann-Gae Gueheneuc and Giuliano Antoniol, "Trustrace: Mining Software Repositories to Improve the Accuracy of Requirement Traceability Links", 2013, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. 5, pp.

AUTHORS PROFILE

Mr. Yogesh Kumara Jakhar is a research scholar in department of Computer Science & Engineering at **Poornima** University, Jaipur. He did his M. Tech (Software Engineering) from Suresh Gyan Vihar University, Jaipur in 2011. Mr. Jakhar has published 10 research papers in various international journals and conferences. His current research interests are data mining, machine learning & multi model ensemble.

Dr. Nidhi Mishra is an associate professor in department of Computer Science & Engineering at Poornima University, Jaipur. She did her PhD in Computer Science & Engineering from Banasthali University Jaipur in 2016. She has published 27 research papers in various international journals and conferences. Her current research interests are Natural language Processing, data mining, machine learning &data science.

Dr. Rakesh Poonia is presently working as Assistant Professor at Govt. Engineering College, Bikaner. Dr. Poonia has contributed many research papers in national and international Journals and conferences. He was member of editorial board at various renowned journals. He was an academic writer with over 14 years of academic research, writing, and editing experience.